BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 8 2007, pages 933–941
doi:10.1093/bioinformatics/btm055

Sequence analysis

Position dependencies in transcription factor binding sites

Andrija Tomovic and Edward J. Oakeley*

Friedrich Miescher Institute for Biomedical Research, Novartis Research Foundation, Maulbeerstrasse 66, CH-4058 Basel, Switzerland

Received on October 29, 2006; revised on January 17, 2007; accepted on February 9, 2007

Advance Access publication February 18, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Most of the available tools for transcription factor binding site prediction are based on methods which assume no sequence dependence between the binding site base positions. Our primary objective was to investigate the statistical basis for either a claim of dependence or independence, to determine whether such a claim is generally true, and to use the resulting data to develop improved scoring functions for binding-site prediction.

Results: Using three statistical tests, we analyzed the number of binding sites showing dependent positions. We analyzed transcription factor–DNA crystal structures for evidence of position dependence. Our final conclusions were that some factors show evidence of dependencies whereas others do not. We observed that the conformational energy (Z-score) of the transcription factor–DNA complexes was lower (better) for sequences that showed dependency than for those that did not ($P < 0.02$). We suggest that where evidence exists for dependencies, these should be modeled to improve binding-site predictions. However, when no significant dependency is found, this correction should be omitted. This may be done by converting any existing scoring function which assumes independence into a form which includes a dependency correction. We present an example of such an algorithm and its implementation as a web tool.

Availability: http://promoterplot.fmi.ch/cgi-bin/dep.html

Contact: edward.oakeley@fmi.ch

Supplementary information: Supplementary data (1, 2, 3, 4, 5, 6, 7 and 8) are available at *Bioinformatics* online.

1 INTRODUCTION

The transcription of genes is controlled by transcription factor proteins (TFs) which bind to short DNA sequences known as transcription factor binding sites (also known as DNA-binding motifs or *cis*-regulatory sequences). TF-binding sites are usually very short and highly degenerate, and such short sequences are expected to occur at random every few hundred base pairs. This makes their prediction extremely difficult. An important task in the computational prediction of TF-binding sites is reducing the false positive rate while still retaining a high sensitivity. Currently, predictions rely on either scanning or *ab initio* methods. Scanning methods infer binding sites from known, experimentally verified binding sequences. Example tools include ConSite (Sandelin *et al.*, 2004a), Match (Kel *et al.*, 2003), Mapper (Marinescu *et al.*, 2005), Patser (Hertz *et al.*, 1990), and rVista (Loots and Ovcharenko, 2004; Loots *et al.*, 2002). *Ab initio* approaches infer specificities without any prior knowledge of binding sites, based on sequence homology. Example tools include Gibbs sampler (Lawrence *et al.*, 1993), MEME (Bailey and Elkan, 1994), Bioprospector (Liu *et al.*, 2001), Yeast motif finder (Sinha and Tompa, 2003) and ANN-Spec (Workman and Stormo, 2000). Until recently, the most popular way of modeling binding sites was to assume that each base in the site occurs independently, e.g. consensus sequence (Day and McMorris, 1992), matrix profiles (Stormo *et al.*, 1982) and sequence logos (Schneider and Stephens, 1990); for a review see (Wasserman and Sandelin, 2004). Methods based on the assumption of independence between positions are simple with small numbers of parameters, making them easy to implement. These methods are widely used and often considered as acceptable models for binding-site predictions (Benos *et al.*, 2002a). However, recent experimental evidence (Benos *et al.*, 2002b; Bulyk *et al.*, 2002; Man and Stormo, 2001; Udalova *et al.*, 2002; Wolfe *et al.*, 1999) has prompted the development of models which incorporate position dependencies. The related methods include Bayesian networks (Barash, 2003), permuted Markov models (Zhao *et al.*, 2005), Markov chain optimization (Ellrott *et al.*, 2002), hidden Markov models (Marinescu *et al.*, 2005), non-parametric models (King and Roth, 2003), and generalized weight matrix models (Zhou and Liu, 2004). Methods based on position-dependency models usually have better binding site prediction accuracy with lower false positive rates. But these methods require more complicated mathematical tools, with more parameters to estimate, and require more experimental data than are typically available (Barash, 2003; Ellrott *et al.*, 2002; King and Roth, 2003; Marinescu *et al.*, 2005; Zhao *et al.*, 2005; Zhou and Liu, 2004). The purpose of this work is to investigate whether or not TFs show position dependencies in their binding sites. We suggest a rigorous statistical approach for testing dependencies. Our findings indicate that there is no universal answer. Some factors seem to show dependencies whereas others do not. We, therefore, decided to allow both possibilities within our model. Our method for modeling dependencies is simply an extension of methods which assume position independencies. It does not require complex

*To whom correspondence should be addressed.

mathematical tools or training data sets (and thus more data), and we will show that it performs much better than existing tools when dependencies are found and no worse when they are not. We also analyzed available structural data to see if any of the observed position dependencies can be explained by 3D structures. We found that dependencies may be partially explained by the 3D structure of TF–DNA complexes. TFs with dependent positions also appear to fit their target sequences better than those without dependencies.

2 METHODS

2.1 Testing dependencies

In this section, we describe methods to test dependencies in binding sites.

Let us suppose that we have n binding sites of length k for a given TF:

$$\begin{matrix} b_1^1 & b_2^1 & \dots & b_k^1 \\ \dots & & & \\ b_1^n & b_2^n & \dots & b_k^n \end{matrix} \tag{1}$$

where $b_i^j \in \{a,c,g,t\}$, and $1 \le i \le k$, and $1 \le j \le n$. We introduce the notation: B_i and B_j to represent random variables which can take values from the set $\{a,c,g,t\}$, indices i and j represent positions in the binding sites and $1 \le i, j \le k$ and $i \ne j$,

$$B_i: \begin{pmatrix} a & c & g & t \\ P(a,i) & P(c,i) & P(g,i) & P(t,i) \end{pmatrix} \tag{2}$$

and likewise for B_j.

Let $N(i)$ be a vector of the frequencies $N(i) = [N(a,i), N(c,i), N(g,i), N(t,i)]$ where, $N(a,i)$ is the frequency of base a at position i and so on. Similarly, for column j we introduce a frequency vector $N(j)$. Using a maximum likelihood approach and the method of Lagrange multipliers, we can estimate probabilities:

$$P(b,i) = \frac{N(b,i)}{n}, \quad P(b,j) = \frac{N(b,j)}{n} \tag{3}$$

where b is one of the bases $\{a,c,g,t\}$.

First, we can calculate mutual information (Chiu and Kolodziejczak, 1991), a quantitative measure of pairwise sequence covariation. The mutual information between positions i and j is given by:

$$M_{ij} = \sum_{b_i,b_j} P(b_i,b_j,i,j) \log_2 \frac{P(b_i,b_j,i,j)}{P(b_i,i)P(b_j,j)} \tag{4}$$

where, the probability $P(b_i,b_j,i,j)$ can be estimated using the maximum-likelihood method and the method of Lagrange multipliers:

$$P(b_i,b_j,i,j) = \frac{N(b_i,b_j,i,j)}{n} \tag{5}$$

where, $N(b_i,b_j,i,j)$ is the frequency of base pairs b_ib_j at positions i and j. This is a descriptive measure of divergence from independence of i and j. M_{ij} varies between 0 and 2 bits. It is maximal when i and j are perfectly correlated. If i and j are uncorrelated, the mutual information is zero. Very often we do not have extreme values of M_{ij}, and we cannot deduce if i and j are independent using only the value of M_{ij}. In order to identify positions that may not be highly correlated as measured by M_{ij}, but are as correlated as they can be given the limited variability of the individual positions, we can calculate two other values (Gutell *et al.*, 1992):

$$R_1(i,j) = \frac{M_{ij}}{H_i}, \quad R_2(i,j) = \frac{M_{ij}}{H_j} \tag{6}$$

where H_i and H_j are entropies for positions i and j, respectively.

$$H_i = -\sum_b P(b,i)\log_2 P(b,i), \quad H_j = -\sum_b P(b,j)\log_2 P(b,j) \tag{7}$$

Both R values vary between 0 and 1 and, in general, they are not equal. Therefore, if we use only M_{ij} we may miss some correlated positions, but some of these may be detected using R-values. However, it should be emphasized that we cannot easily estimate the significance of R-values. So, we will have false positives as well as true correlated positions. R-values are also descriptive measures of dependencies between two positions. A more formal way to test dependencies is hypothesis testing:

$$\begin{aligned} &H_0\text{: positions } i \text{ and } j \text{ are independent} \\ &H_1\text{: otherwise.} \end{aligned} \tag{8}$$

To test this hypothesis, we can use a χ^2-test of independence (Ellrott *et al.*, 2002) on each pair of positions i and j:

$$X^2 = \sum_{b_i,b_j} \frac{(P(b_i,b_j,i,j) - P(b_i,i)P(b_j,j))^2}{P(b_i,i)P(b_j,j)} \tag{9}$$

The distribution of X^2 statistics is close to a χ^2 distribution with $(|b_i|-1)\times(|b_j|-1)$ degrees of freedom, where $|b_i|$ is the number of bases for which $P(b_i,i)$ is not zero, and likewise for $|b_j|$. So, using X^2 statistics and χ^2 distributions we can test the hypothesis at a given significance level e.g. 0.05. This hypothesis may also be tested using a G-test of independence (log-likelihood ratio test) (Sokal and Rohlf, 2003). For each pair of positions i and j, we can calculate G statistics:

$$G = 2\sum_{b_i,b_j} P(b_i,b_j,i,j) \ln\left(\frac{P(b_i,b_j,i,j)}{P(b_i,i)P(b_j,j)}\right) \tag{10}$$

The distribution of G statistics is close to χ^2 with $(|b_i|-1)\times(|b_j|-1)$ degrees of freedom where $|b_i|$ is the number of bases for which $P(b_i,i)$ is not zero, and likewise for $|b_j|$. M_{ij} corresponds to a G-statistics value if we log transform it. A general problem with both χ^2 and G-tests is small sample sizes, i.e. small expected frequencies (in our notation these are the values $nP(b_i,i)$ and $nP(b_j,j)$). This is because the number of known binding sites is usually small. Cochran (Cochran, 1954) suggested that independence may be tested so long as we have more than one degree of freedom. A minimum expected value of 1 is allowed, provided that no more than 20% of the categories have expected values below 5. Here, X^2 statistics have been shown to be valid with fewer samples and more sparse tables than G statistics. The G-statistic distribution is usually a poor approximation to χ^2 when expected frequencies are <5 (Agresti, 1990; Koehler, 1986; Koehler and Larntz, 1980; Larntz, 1978). William's correction for G (Williams, 1976) partially addresses this:

$$G_{\text{adj}} = \frac{G}{q}, \quad q = 1 + \frac{(a^2-1)}{6nv} \tag{11}$$

where, $a = (|b_i|-1)\times(|b_j|-1) - 1$, and $v = a - 1$ as this provides a better approximation to the χ^2 distribution. Conahan found that if expected frequencies are higher than 10, G statistics approximate well to the exact multinomial probability distribution (Conahan, 1970). She found that G statistics were adequate and better than X^2 statistics, where there are more than five degrees of freedom and expected frequencies greater than or equal to 3. In all other cases she recommends the exact test. Larantz, in his comparison of G and X^2 statistics, did not consider the corrections of G statistics when drawing his conclusion that X^2 statistics fits the theoretical chi-squared distribution better than G statistics do (Larntz, 1978). Sokal *et al.* (Sokal and Rohlf, 2003) showed that G statistics with William's correction approximates to the χ^2 distribution more closely than they do without the correction. It is very difficult to find a single rule to cover all cases when the observed distributions of G statistics and X^2 statistics are close to real χ^2 distributions, if we have small expected

frequencies (Agresti, 1990). A safer way to test the hypothesis of dependence is, therefore, to use exact methods like the exact randomization (nonparametric) test (Sokal and Rohlf, 2003). The problem with this test is that, even though we have small sample numbers, there are a large number of possible outcomes, and their complete enumeration is impractical. Because of this, we have to use a Monte Carlo randomization test (Davison and Hinkley, 1997; Manly, 1997), in which the problem is solved by random sampling from a simulated population. Monte Carlo randomization tests can be performed using X^2 or G statistics. We used X^2 statistics with 10 000 replications in the statistics package R (GNU software).

Two random variables b_i and b_j are independent if

$$P(B_i, B_j) = P(B_i)P(B_j). \tag{12}$$

Thus we can test the following hypotheses for dependence/independence (instead of hypothesis testing (9)):

$$\begin{aligned} &H_0\text{: distributions } P(B_i, B_j) \text{ and } P(B_i)P(B_j) \text{ are the same} \\ &H_1\text{: otherwise.} \end{aligned} \tag{13}$$

This form of hypothesis testing corresponds to a multinomial goodness-of-fit test. As in (Bejerano, 2003, 2006; Bejerano *et al.*, 2004), we can test for a correlation between TF-binding site positions using exact P-values (for hypothesis testing (13)). This approach gives more accurate results than either χ^2 or G-tests (Bejerano, 2003, 2006; Bejerano *et al.*, 2004). The only problem with this approach is that it is computationally expensive. However, a recent publication (Keich and Nagarajan, 2006) has shown that grid approximations yield almost identical results for the P-values but in far less time (Bejerano, 2006). The final method we have used to test dependencies is a Bayesian approach (Minka, 2003; Zhou and Liu, 2004). We can calculate the Bayes factor $BF(H_0; H_1)$ for hypothesis testing as follows (full derivation of formula (4) can be found in Supplemental Material 1—derivation 1)

$$\begin{aligned} BF(H_0; H_1) = &\frac{\Gamma\left(\sum_{b_i, b_j} \alpha_{b_i b_j}\right)}{\Gamma\left(n + \sum_{b_i, b_j} \alpha_{b_i b_j}\right)} \prod_{b_i} \frac{\Gamma\left(N(b_i, i) + \alpha_{b_i}\right)}{\Gamma\left(\alpha_{b_i}\right)} \\ &* \prod_{b_j} \frac{\Gamma\left(N(b_j, j) + \alpha_{b_j}\right)}{\Gamma\left(\alpha_{b_j}\right)} \prod_{b_i, b_j} \frac{\Gamma\left(\alpha_{b_i b_j}\right)}{\Gamma\left(N(b_i, b_j, i, j) + \alpha_{b_i b_j}\right)} \end{aligned} \tag{14}$$

We choose $\alpha_{b_i b_j} = 1$ and $\alpha_{b_i} = \sum_{b_j} \alpha_{b_i b_j}$ and the calculation should include only bases b_i, b_j for which $N(b_i, i) \neq 0$ and $N(b_j, j) \neq 0$.

Using Stirling's approximation ($\log\Gamma(x+1) \approx x \log x - x$) it can be shown that (Supplemental Material 1—derivation 2)

$$\log_2(BF(H_0; H_1)) \approx -nM_{ij} \tag{15}$$

This gives us the relationship between BF and mutual information (Minka, 2003). The relationship between these two values is better when the sample size n is higher (due to the use of Stirling's approximation). We used formula (20) to calculate BF, and report that when $BF(H_0; H_1) < 0.1$ there is strong evidence against the null hypothesis.

Thus, in this article we used three distinct methods for dependence testing between the TF site base positions. These methods were:

(i) Monte Carlo randomization test with X^2 or G statistics
(ii) Exact multinomial goodness-of-fit test
(iii) Bayesian hypothesis testing.

There is always a danger of type I errors (rejecting the null hypothesis when in fact it is true) when applying multiple tests to data. These may be minimized with Bonferroni's correction or its extensions/variants (e.g. Dunn–Šidák, Holm's, Simes–Hochberg or Hommel's method). The Bonferroni adjustment of P-value ($0.05/k$, where k is the number of tests) is very stringent and can introduce type II errors, which are also important. The use of Bonferroni is much debated (Perneger, 1998).

As a compromise, in the case of the Bayesian test, we propose that a more stringent BF factor $BF(H_0; H_1) < 0.1$ could be used to report stronger evidence against the null hypothesis.

2.2 New scoring function

Any existing scoring function which works with models that assume independence between positions within binding sites, can easily be modified to incorporate dependencies. These new functions do not have dramatically more parameters, and do not require additional data or complex mathematical approaches.

If we have n binding sites of length k for a given TF and sequence l with length k, then to determine if a putative-binding site is for a given TF we will follow the notation of (Wasserman and Sandelin, 2004) where, $w_{b,i}$ is a position weight matrix (PWM) value of base b in position i, calculated by:

$$W_{b,i} = \log_2 \frac{P(b, i)}{P(b)} \tag{16}$$

where $P(b)$ is the background probability of base b ($P(b) = 0.25$) and $P(b, i)$ is a corrected probability of base b at position i, and is calculated by:

$$P(b, i) = \frac{N(b, i)}{n} + a(b) \tag{17}$$

where $a(b)$ is smoothing parameter ($a(b) = 0.01$).

The fit of any given DNA sequence can be quantitatively scored by summing all the values of $W_{b,i}$ for every base in the sequence (hereafter, we will refer to this 'old' scoring function as S_{old}):

$$S_{\text{old}} = \sum_{i=l}^{k} w_{l_i, i} \tag{18}$$

For a large set of well-characterized binding sites, these scores are proportional to the factor-binding energies (Stormo, 2000).

To incorporate position dependencies, we will extend this function and this model for the representation of the TF-binding sites in the following way.

First, we will introduce a corrected probability for the bases $b_1 b_2 \ldots b_m$ in $i_1 i_2 \ldots i_m$ dependent positions.

$$P(b_1, \ldots, b_m, i_1, \ldots, i_m) = \frac{N(b_1, \ldots, b_m, i_1, \ldots, i_m)}{n} + a(b_1, \ldots, b_m) \tag{19}$$

$a(b_1, \ldots, b_m)$ is a smoothing parameter and can be calculated by:

$$a(b_1, b_2, \ldots, b_m) = a(b_1) \ldots a(b_m) \tag{20}$$

Then we can calculate values which correspond to PWM values:

$$W_{b_1, \ldots, b_m, i_1, \ldots, i_m} = \log_2 \frac{P(b_1, \ldots, b_m, i_1, \ldots, i_m)}{P(b_1) \ldots P(b_m)} \tag{21}$$

Finally, the new scoring function (S_{new}), which incorporates dependencies, can be expressed thus:

$$\begin{aligned} S_{\text{new}} = &\sum_{i=l}^{k_1} W_{l_i, i} + \sum_{i=1}^{k_2} W_{l_{j_i}, l_{j_{i+1}}, j_i, j_{i+1}} + \cdots + \\ &+ \sum_{i=l}^{k_m} W_{l_{j_i}, \ldots, l_{j_{i+m-1}}, j_i, \ldots, j_{i+m-1}} \end{aligned} \tag{22}$$

where, k_1 is the number of independent positions, k_2 is the number of dependent positions order 2 (nucleotides at positions j_i and j_{i+1}) and k_m the number of dependent positions order m (nucleotides at positions j_i, $j_{i+1}, \ldots, j_{i+m-1}$). Higher-order dependencies can be constructed from

the second-order dependencies in the following way: if we analyze three positions i_1, i_2 and i_3, and if every two combinations ($i_1 - i_2$, $i_1 - i_3$ and $i_2 - i_3$) are dependent, then we can claim that positions i_1, i_2 and i_3 show third-order dependencies. This approach may be extended to find mth-order dependencies between k_m positions of a binding site. For the new scoring function (22), higher order dependencies can be constructed in a less stringent way: if we find when analyzing three positions i_1, i_2 and i_3 that only two combinations ($i_1 - i_2$, $i_2 - i_3$ or $i_1 - i_3$) are dependent, we can say that there are third order dependencies among positions i_1, i_2 and i_3. This will not have any influence on the final results (because of equation (12)) and the logarithm property ($\log(P(B_i, B_j))$ will tend towards $\log(P(B_i)) + \log(P(B_j))$). Small differences may be observed because of the smoothing parameters, but this helps in the practical implementation of new scoring function.

Binding scores calculated by the scoring functions S_{old} and S_{new} can be normalized according to (Bucher, 1990; Tsunoda and Takagi, 1999):

$$S'_{\text{old}} = \frac{S_{\text{old}} - S^{\min}_{\text{old}}}{S^{\max}_{\text{old}} - S^{\min}_{\text{old}}}, \quad S'_{\text{new}} = \frac{S_{\text{new}} - S^{\min}_{\text{new}}}{S^{\max}_{\text{new}} - S^{\min}_{\text{new}}} \tag{23}$$

where $S^{\min}_{\text{old}}$, $S^{\max}_{\text{old}}$ are the hypothetical minimum and maximum for S_{old} and $S^{\min}_{\text{new}}$, $S^{\max}_{\text{new}}$ are the hypothetical minimum and maximum for S_{new} (analytic formula for their calculation is given in Supplemental Material 1).

For the final implementation of the function (22), it is useful to construct sequence dependency corrected matrices of TFs. However, in practice, this can be very inefficient because the dimensions of these matrices can be very high with a lot of zeros. Because of this, we provide a database (available at http://www.fmi.ch/members/andrija.tomovic/database.txt) with sequences and dependent positions written below (estimated using a Monte Carlo randomization test with X^2 without Bonferroni's correction or exact multinomial goodness-of-fit without Bonferroni's correction or Bayesian hypothesis testing with $BF(H_0; H_1) < 0.1$ and higher order of dependencies in less stringent variant). This is a compact and readable format of sequence dependency corrected matrices of TFs from the JASPAR database (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004b). For the identification of TF-binding sites by scoring function (22), we suggest using the all-atom model, like it is used with function (18). In combination with databases of good quality binding sites (such as JASPAR) all-atom methods give better accuracy. If we cut the length of binding sites, there may be dependent positions in this region which will be lost to our function (22). Both the new (22) and old (18) scoring functions are linear in complexity, so cutting would not improve performance much.

3 RESULTS AND DISCUSSION

3.1 Distributions of transcription factors with dependent positions

To determine the distributions of TFs with dependent positions, we used the public database JASPAR (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004b) which contains experimentally determined, high-quality binding sites. The JASPAR database represents a curated and non-redundat data-set (Lenhard and Wasserman, 2002; Sandelin *et al.*, 2004b). We selected all TFs for which there were binding sequences (not only matrix profiles) and the final data set contained 107 TFs with 3239 binding sites. We applied three different tests (Section 2.1) to each of these binding sites to establish how many factors showed position dependencies (Table 1). We also show the effect of either applying Bonferroni's corrections or using the more stringent $BF(H_0; H_1) < 0.1$ cut off. Rows A, B and C of Table 1 may include some false positives, but rows D, E and F have a false negative problem. A complete list of every pair of positions for each TF is given in Supplemental Material 2. We also report values of M_{ij}, R_1 and R_2, as well as G-statistic values with their degrees of freedom and P-values. In addition, we report the adjusted G-statistic values with their degrees of freedom and adjusted G-test P-values; the X^2 statistics together with their degrees of freedom and P-values; and also the average value of expected frequencies and the percentage of expected values smaller than 5 and smaller than 3. Finally, in this table we report the P-values of the Monte Carlo randomization test with X^2 statistics, the exact multinomial 'goodness-of-fit' test and the Bayesian factor (BF) values. From this analysis, we observe that the sample sizes are not appropriate for either chi-squared or G-tests of independence (column H in Supplemental Material 2). As discussed previously (Section 2.1), this implies that these two tests will give poor probability estimates. The values of M_{ij}, R_1 and R_2 may be used as descriptive measures of position associations. There is good agreement between results produced using the three 'statistically correct' tests we attempted. The most stringent is the exact multinomial goodness-of-fit test, and the least stringent is the Monte Carlo randomization test. Almost every pair of dependent positions predicted by the exact multinomial goodness-of-fit test is also reported by the other two tests. The Monte Carlo randomization test gives more precise probabilities than either the chi-squared or G-tests, but with low power because of the lack of experimental data (small sample size).

Table 1. Distributions of TFs with dependent positions tested

Statistical test	TFs with dependent positions
A	74.77%
B	49.52%
C	62.62%
D	38.32%
E	23.26%
F	26.17%

A—Monte Carlo randomization test without Bonferroni's correction; B—Exact multinomial 'goodness-of-fit' test without Bonferroni's correction; C—Bayesian hypothesis testing $BF(H_0; H_1) < 0.1$; D—Monte Carlo randomization test with Bonferroni's correction; E—Exact multinomial 'goodness-of-fit' test with Bonferroni's correction; F—Bayesian hypothesis testing $BF(H_0; H_1) < 0.01$.

In addition, we looked to see if the length and number of known binding sites were different between the groups of TFs with and without dependent positions (Table 2). The variances of these two groups are not statistically different (tested by Bartlett's test). Using Student's t-test, we tested the null hypothesis that mean length and number of binding sites between the two groups are equal against a one-tailed alternative hypothesis that TFs without dependent positions have shorter lengths and smaller numbers of known binding sites. In each case, we obtained P-values less than 0.05 and thus we should reject the null hypothesis and accept the alternative.

Table 2. Average length and number of binding sites between a group of TFs with dependent positions and a group of TFs without dependent positions

Statistical test	Average length of TFs binding sites		Average number of known binding sites	
	I	II	I	II
A	11.67	8.25	32.85	22.64
B	12.15	9.43	34.66	25.77
C	11.66	9.3	35.791	20.775
D	12.19	9.89	39.15	24.61
E	11.92	10.265	45.04	25.82
F	12.00	10.34	50.96	22.91

I—group with dependent positions; II—group without dependent positions; A, B, C, D, E, F — notation the same as in Table 1.

These results imply that more factors may show dependencies once additional binding-site data becomes available.

Based on the second-order dependencies (dinucleotide dependencies), it is possible to construct higher order dependencies, as explained in section 2. It is clear that when we have such dependencies, there are lower order dependencies in all combinations. Because of this, it is useful to analyze distributions of dependencies of different orders k_m ($2 \leq k_m \leq 9$) constructed in a more stringent way (Supplemental Material 3). We analyzed the distributions of TFs with dependent positions in structural classes of TF–DNA-binding domains. We wanted to investigate whether there is any tendency for certain folds to have position dependencies (Supplemental Material 4). We noticed that some structural classes contain TFs with position dependencies in their binding sites detected by almost all statistical tests, such as: T-BOX, P53, AP2, TRP, CAAT-box and MADS. Other classes contain TFs without dependent positions like: ZH-FINGER-DOF, ZH-FINGER-GATA, HOMEO/CAAT and 'Unknown' class. However, the major structural classes contain TFs with and without dependent positions (bZIP, nuclear receptor, etc.).

3.2 Do position dependencies relate to 3D structures?

We wanted to investigate possible biological explanations of the dependent positions we predicted. We investigated this by examining 3D crystal structures when available. Possible explanations of dependency include:

- active amino acids might interact with dependent nucleotides either singly or in pairs via hydrogen bonds or salt bridges;
- conformational changes in the structure of DNA caused by one dependent base may alter the accessibility of the other dependent bases to the binding site;
- something else.

We selected 32 TF–DNA co-crystal pairs of structures from the PDB database at resolutions better than 3.0 Å (Berman *et al*., 2000) corresponding to TFs with published binding sites in JASPAR (September 2006) (Table Sup3-1 in Supplemental Material 3). Direct contacts between bases and amino acids were investigated (Table Sup3-2 in Supplemental Material 3).

There is no clear one-to-one correspondence between dependent DNA-binding positions and their interactions with TF. This is not a big surprise because these proteins recognize specific DNA sequences not only via direct contact but also indirectly, through specific sequence-dependent DNA conformations, distortions or water-mediated contacts (Sarai and Kono, 2005). Amino acids neighboring dependent bases may be different from those around independent positions. In addition, mutations in bases which do not directly contact the amino acid may still affect the binding affinity (see references listed in Sarai and Kono, 2005).

Next, we wanted to check whether there were any relationships between dependent positions and conformational changes of the DNA. We could calculate structural parameters to describe the 3D nucleic acid structures using the software package 3DNA (Lu and Olson, 2003), but there are many parameters (shift, slide, rise, tilt, roll and twist) to describe the structure of DNA, and because we have relatively few sequences in our data set it is difficult to identify significant effects. Similarly, if we want to investigate spatial distribution patterns of neighboring amino acids around dependent positions, we will have a data-mining problem.

We decided to use the energy Z-scores (Ahmad *et al*., 2006; Gromiha *et al*., 2004; Kono and Sarai, 1999) for TF–DNA complexes for both 'direct' and 'indirect' readouts. The energy Z-score for direct readouts quantifies the spatial distributions of side chains around base pairs, and represents the base–amino acid interaction energy. The energy Z-score for indirect readouts quantifies DNA conformation, and represents the conformational energy of DNA. The more negative the Z-score, the better the target sequence fits into a given structure (Ahmad *et al*., 2006). The list of all Z-score values can be found in Supplemental Material 4. We tested the Z-scores using a one-tailed Student's *t*-test (Table 3). The direct readout showed no difference between TFs with dependent or independent positions ($P > 0.1$). However, the conformational energy (indirect readout) was always significantly lower for TFs with dependent positions ($P < 0.02$). This means that TFs with dependent positions fit their target DNA motifs better than those without. These results suggest a possible relationship between position dependencies and the 3D structure of TFs.

Table 3. Average Z-score for direct and indirect readout for: **I**—a group of TFs with dependent positions; and **II**—a group of TFs without dependent positions

Statistical test	Average Z-score (direct readout)			Average Z-score (indirect readout)		
	I	II	*P*-value	I	II	*P*-value
A	−2.5	−2.62	–	−2.8	−1.791	0.00565**
B	−2.67	−2.42	0.383	−3.0914	−2.01	0.0016**
C	−2.667	−2.25	0.31	−2.747	−1.907	0.02*
D	−3.054	−2.26	0.17	−3.22	−2.09	0.00152**
E	−3.44	−2.3	0.111	−3.33	−2.29	0.0147*
F	−3.1025	−2.32	0.186	−3.497	−2.147	0.0005***

*$P<0.05$, **$P<0.01$, ***$P<0.001$.
A, B, C, D, E, F—notation the same as in Table 1. The variances of groups I and II are not statistically different (Bartlett's test).

We investigated if DNA sequence length influences the conformational energy. In the 32 cases we studied where we have both a 3D crystal structure and a JASPAR matrix ID, we performed one- and two-tailed *t*-tests on the lengths of sequences found to show dependencies and those without dependencies from each of the six dependency tests we investigated. These results showed that five out of six of the tests (two-tailed) and three out of six (one-tailed) do not show significant differences in sequence length between the two groups for which we have conformational energies (Supplemental Material 6). If the conformation of the DNA fragment is not sequence specific, then the conformational energy is expected to fluctuate independently of fragment size. But, if the conformation is sequence specific, then the total energy should decrease with the size although the average energy per base will not decrease if the energy distribution is uniform (A. Sarai, personal communication). For these reasons, we believe that sequence length is not the major factor contributing to the significantly lower conformational energies we found for the group of TFs with dependent positions.

We analyzed relationships between dependent position and DNA stiffness to show the influence of DNA stiffness on protein–DNA binding specificity (Gromiha, 2005). We calculated the average stiffness of DNA using the structure-based sequence-dependent stiffness scale (Gromiha, 2005) for binding sites with and without position dependencies (Supplemental Material 7). In two cases, we found that the average stiffness values are significantly larger (one-tailed Student's *t*-test $P<0.028$) for sites with dependent positions (detected by Bayesian hypothesis testing in both variants) than without dependent positions. However, in the other four cases no significant differences were found.

3.3 Evaluation of a new scoring function for the prediction of TF-binding sites

The evaluation of *ab initio* methods for the prediction of TF-binding sites is described in (Tompa *et al.*, 2005). Here, we will perform a slightly different validation. In order to evaluate the new scoring function given by (22) and (23), we performed a validation using both synthetic and experimentally verified data.

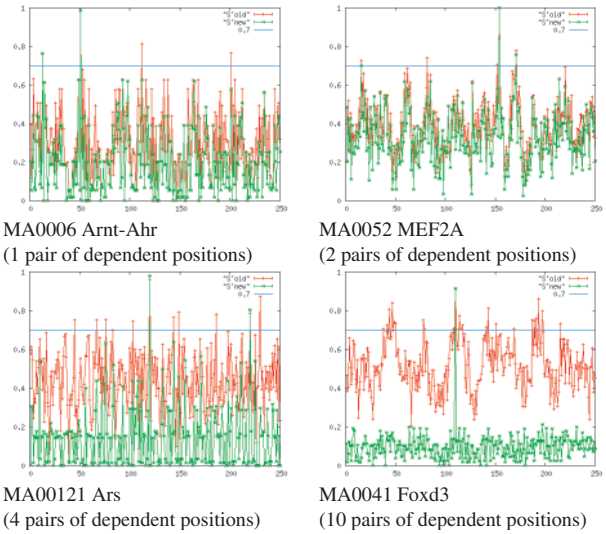

Fig. 1. Comparison of old and new scoring functions with synthetic data.

First, we generated a random sequence from a third-order Markov model background distribution using the program RSA (van Helden, 2003). In this sequence, we planted binding site 9 of the TF MA0006 at position 51. We had found one dependent position in this TF. We then calculated a normalized scoring value for each position in the sequence, using both the old and new functions. We assigned a threshold of 0.7 as indicating a match for a binding site (Fig. 1). The new scoring function made one false-positive prediction and one true positive, whereas the old scoring function made three false-positive predictions and one true positive. We repeated this with similar experiments (data available at http://www.fmi.ch/members/andrija.tomovic/exp1.zip) using: MA0052 (two pairs of dependent positions); MA0121 (four pairs of dependent positions); and MA0041 (10 pairs of dependent positions). The accuracy of the new scoring function improved as the number of dependent positions increased. The so-called 'twilight zone' region of the plots also becomes narrower with a smaller density. If there are no dependent positions, then the new and

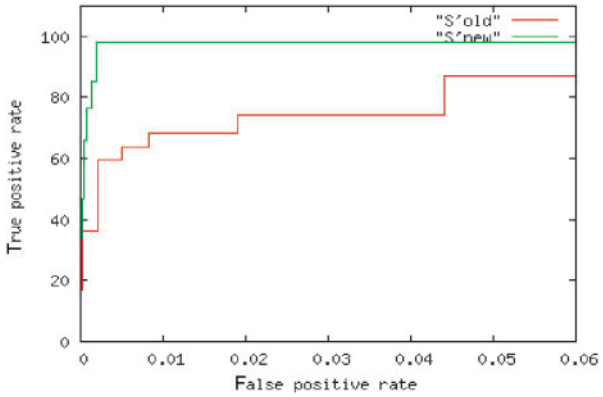

Fig. 2. ROC curves for new and old scoring functions, showing their ability to predict binding sites. The *x*-axis shows the false-positive rate $(FP/(FP+TN)) \times 100$, the *y*-axis shows the true-positive rate $(TP/(TP+FN)) \times 100$.

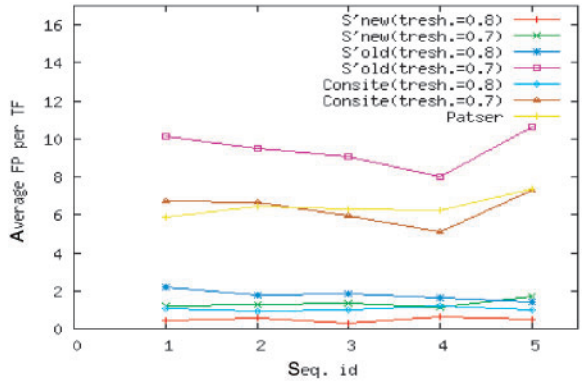

Fig. 3. Average false-positive ratio per TF for different prediction methods.

old scoring functions are the same. We currently only apply our correction for positions that show statistically significant dependencies. If, instead, we factor the observed frequency *P*-scores for all bases, regardless of their significance, then the new function will tend towards the old function because of Equation (12), and the logarithm property ($\log(P(B_i, B_j))$ will tend towards $\log(P(B_i)) + \log(P(B_j))$ but small differences may be observed because of the smoothing parameters. The price for doing this is computational time, and it does not appear to offer any great advantage over the solution we have implemented.

To further evaluate our new scoring function, we generated 1850 random sequences sampled from a third-order Markov model background distribution with lengths from 250 to 500. In 50, we planted binding sites for MA0041 Foxd3, and we then analyzed the true- and false-positive rates for different threshold values using the new and old scoring functions (Fig. 2 and Table Sup8-1 in Supplemental Material 8). Both functions have good scores for true positives, but the new scoring function gave better results. The biggest difference was in the false-positive rate which was much better with the new scoring function. Next, we generated five random sequences sampled from a third-order Markov model background distribution (with lengths from 400 to 600) in which we planted 0–3 binding sites from a set of 15 (all 15 contained dependent positions). The data set is given in Supplemental Material 8. We wanted to measure the accuracy of prediction with the new scoring function and compare it with other available tools and methods (PATSER, ConSite and the old scoring function). Given that almost all of the methods can detect true positives (i.e. they have a high sensitivity), the accuracy of each method should be estimated by its selectivity (false-positive rate). These results are shown in Figure 3 and Table Sup8-2 in Supplemental Material 8. Our new scoring function (22-23) performed best with the smallest number of false positives per nucleotide and per TF. Finally, we analyzed real experimental data. As ConSite had the next best prediction results with the synthetic data, we decided to use it for benchmark comparisons with the experimental data as well. We used a set of genes showing skeletal muscle-specific expression (Wasserman and Fickett, 1998). This set is an updated version from (Defrance and Touzet, 2006) which has been used to evaluate such tools in the past. This dataset includes upstream regions (2000 bp) of nine genes (see Table Sup8-3 in Supplemental Material 8) and six TFs from the JAPSAR database (MA0052, MA0055, MA0056, MA0057, MA0079 and MA0083) which are known to be involved in the regulation of skeletal muscle-specific expression. MA0055's binding sites are not listed in JASPAR, so its detection will be unchanged from the old function (24). We scanned the upstream sequence of the nine genes using all of the TFs from JASPAR. There are 16 TFs (including MA0055) for which there is no binding sequence information, only weight matrices. These will be treated as having independent binding (24), which will have a negative effect on the results from the new scoring function, but is more realistic. However, even with this limitation, the results from the new scoring function are slightly better than those from ConSite (TableSup 8-3 in Supplemental Material 8). The false-positive rate for all nine sequences is smaller with the new scoring function, and the true-positive rate is almost the same. ConSite detected one true positive hit more (for three sequences) than our scoring function with this data set.

4 CONCLUSIONS

In this work, we performed a detailed analysis of dependencies within TF-binding sites. Our conclusion is that we cannot assume that positions are either dependent or independent. This must be tested using one of three proposed statistical tests. Our structural analysis indicates that some of the predicted dependencies agree with 3D structure data from TF–DNA complexes. We propose that the dependencies we have identified should be used in binding-site predictions. Previous attempts at such modeling have required complex tools with many parameters which really require more training data than is currently available. Here, we present a simple way of modeling these dependencies. We demonstrated how to modify existing dependence-free scoring functions to consider dependencies. Such modifications improve prediction quality for TFs

with dependent positions. Our technique does not require complex tools or more training data than scoring functions and models which assume independence. This approach can be used with any scoring function which assumes independence (one such is presented here). We demonstrated this approach using scanning methods for the prediction of TF-binding sites, but it can be applied to work with *ab initio* methods and different methods of prediction which incorporate comparative genomic analysis (phylogenetic footprinting conservation).

ACKNOWLEDGEMENTS

We would like to thank Prof Gill Bejerano, Prof Frank Hampel, Dr Hans-Rudolf Roth, Dr Michael Stadler and Prof Akinori Sarai for useful discussions and advice. This work was supported by the Novartis Research Foundation. Funding to pay the Open Access publication charges was provided by the Novartis Research Foundation FMI.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (1990) *Categorical Data Analysis*. John Wiley Sons, New York.

Ahmad,S. *et al*. (2006) ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. *Nucleic Acids Res*., **34**, W124–W127.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **2**, 28–36.

Barash,Y. *et al*. (2003) Modeling dependencies in protein-DNA binding sites. In *Proceedings of RECOMB-03*, 28–37.

Bejerano,G. (2003) Efficient exact p-value computation and applications to biosequence analysis. In *Proceedings of RECOMB-03*, 38–47.

Bejerano,G. (2006) Branch and bound computation of exact p-values. *Bioinformatics*, **22**, 2158–2159.

Bejerano,G. *et al*. (2004) Efficient exact p-value computation for small sample, sparse, and surprising categorical data. *J. Comput. Biol*., **11**, 867–886.

Benos,P.V. *et al*. (2002a) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*., **30**, 4442–4451.

Benos,P.V. *et al*. (2002b) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol*., **323**, 701–727.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*., **212**, 563–578.

Bulyk,M.L. *et al*. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*., **30**, 1255–1261.

Chiu,D.K. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci*., **7**, 347–352.

Cochran,W.G. (1954) Some methods for strengthening the common chi-square tests. *Biometrics*, **10**, 417–451.

Conahan,M.A. (1970) The comparative accuracy of the likelihood ratio and Chi-squared as approximation to the exact multinomial test. Lehigh University, 64.

Davison,A.C. and Hinkley,D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Day,W.H. and McMorris,F.R. (1992) Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res*., **20**, 1093–1099.

Defrance,M. and Touzet,H. (2006) Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics*, **7**, 396.

Ellrott,K. *et al*. (2002) Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, **18** (Suppl. 2), S100–S109.

Gromiha,M.M. (2005) Influence of DNA stiffness in protein-DNA recognition. *J. Biotechnol*., **117**, 137–145.

Gromiha,M.M. *et al*. (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol*., **337**, 285–294.

Gutell,R.R. *et al*. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*., **20**, 5785–5795.

Hertz,G.Z. *et al*. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci*., **6**, 81–92.

Keich,U. and Nagarajan,N. (2006) A fast and numerically robust method for exact multinomial goodness-of-fit test. *J. Comput. Graph. Stat*., **15**, 779–802

Kel,A.E. *et al*. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.

King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*., **31**, e116.

Koehler,K.J. (1986) Goodness-of-fit test for log-linear models in sparse contingency tables. *J. Am. Stat. Assoc*., **81**, 483–493.

Koehler,K.J. and Larntz,K. (1980) An empricial investigation of goodness-of-fit statistics for sparse multinomials. *J. Am. Stat. Assoc*., **75**, 336–344.

Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.

Larntz,K. (1978) Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *J. Am. Stat. Assoc*., **73**, 253–263.

Lawrence,C.E. *et al*. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Liu,X. *et al*. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *In proceedings of Pac. Symp. Biocomput*., 127–138.

Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*., **32**, W217–W221.

Loots,G.G. *et al*. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*., **12**, 832–839.

Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*., **31**, 5108–5121.

Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res*., **29**, 2471–2478.

Manly,B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London.

Marinescu,V.D. *et al*. (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, **6**, 79.

Minka,T. (2003) Bayesian inference, entropy, and the multinomial distribution. Technical Report (Microsoft research).

Perneger,T.V. (1998) What's wrong with Bonferroni adjustments,. *BMJ*, **316**, 1236–1238.

Sandelin,A. *et al*. (2004a) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*., **32**, W249–W252.

Sandelin,A. *et al*. (2004b) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.

Sarai, A. (2006/07) Personal Communication.

Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Ann. Rev. Biophys. Biomol. Struct*., **34**, 379–398.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.

Sinha,S. and Tompa,M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*., **31**, 3586–3588.

Sokal,R.R. and Rohlf,F.J. (2003) *Biometry: The Principle and Practice of Statistics in Biological Research*. W.H. Freeman and Company, New York.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. *et al*. (1982) Characterization of translational initiation sites in E. coli. *Nucleic Acids Res*., **10**, 2971–2996.

Tompa,M. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol*., **23**, 137–144.

Tsunoda,T. and Takagi,T. (1999) Estimating transcription factor bindability on DNA. *Bioinformatics*, **15**, 622–630.

Udalova,I.A. *et al*. (2002) Quantitative prediction of NF-kappa B DNA-protein interactions. *Proc. Natl. Acad. Sci. USA*, **99**, 8167–8172.

van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res*., **31**, 3593–3596.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions \which confer muscle-specific gene expression. *J. Mol. Biol*., **278**, 167–181.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*., **5**, 276–287.

Williams,D.A. (1976) Improved likelihood ratio tests for complete contingency tables. *Biometrika*, **63**, 33–37.

Wolfe,S.A. *et al*. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol*., **285**, 1917–1934.

Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. In *proceedings of Pac. Symp. Biocomput.,* 467–478.

Zhao,X. *et al*. (2005) Finding short DNA motifs using permuted Markov models. *J. Comput. Biol*., **12**, 894–906.

Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.